Wind Turbine Power Output Forecast

by Amit Bharadwa

January 2020

Mentor: Wayne Ang

Contents

1 Problem Statement

There is no doubt that renewable forms of energy are great for the environment, however, they do have their flaws. The biggest problem is intermittency and this ends up being a problem for the national grid trying to meet supply and demand. This report will focus on addressing this problem by using data recorded by a SCADA (Supervisory control and data acquisition) system, and by building a predictive deep learning model to forecast the power production of a wind turbine over two weeks with hourly intervals. Based on these predictions, the model will be able to predict the intermittency of wind and provide insight on how to make optimal delivery commitments to the grid.

2 Introduction

Forms of renewable energy are great for the environment and in the case of wind turbines, zero carbon emission when they're up and running. With wind turbines ranging in size from sitting on a work desk to as large as competing in vertical height with "The Gherkin" in London. It comes as no surprise the immense power these monsters can generate and with an approximate 20 year life span, a vast number of homes can receive a sustainable form of energy for many years. The problem arises with intermittency, the wind is not a reliable source of energy.

With other conditions such as wind speeds need to exceed $\approx 3.3m/s$ for a wind turbine to experience torque, careful consideration needs to be taken for choosing locations to erect a turbine. Even with careful planning and preparation, there are very few locations where a turbine can expect to reach it's full potential. As an unreliable form of energy, it can be difficult for the national grid rely heavily on wind energy. However, if it is possible to predict intermittency, the national grid can prepare for the upcoming events in advance, by meeting the demand with another sources of renewable energy. Solving this problems ensures stability in the national grid whilst minimizing the problem of supply and demand and the overall goal of a sustainable future.

This report will go into detail about a single wind turbine located in Turkey during 2018. Furthermore, a focus on the performance of the turbine over the course of the year, in addition to statistical testing to extract and identify relevant features from the data-set. Lastly, a prediction from a deep learning model which will predict the power output of the turbine over two weeks with hourly intervals.

3 Understanding the Data

The data used for this project was acquired from Kaggle. Originally recorded using a SCADA device, which collects data at the height of the hub of a wind turbine in Turkey. The recorded data covers the year of 2018 in 10 minute intervals and is for a single wind turbine.

3.1 The Data

Prior to modelling the data, a better understanding of the data is required. This is so missing values can be identified, explore the data using visualisation tools and finding periods during the year the turbine is under maintenance. The original data used in this project consists of five columns.

- Date/Time: 10 minute intervals
- LV Active Power (kW): The power generated by the turbine at that moment in time
- Theoretical Power Curve (kWh): The theoretical maximum power the turbine can produce at that moment due to the wind speed.
- Wind Direction (°) : The wind direction at hub height.

3.2 Data Wrangling & Feature Engineering

The original dataset contains 50530 entries with no missing values. The dataset is resampled into hourly intervals, as the model will try to predict the power output every hour and data which meets the criteria $WindSpeed >= 3.3m/s$ & $LV ActivePower <= 0$ are removed from the dataset. These are clear signs the turbine is under maintenance. The remaining missing values are interpolated.

Additional features are appended to the end of the dataset. With statistical testing an analysis can be done on whether the new features are relevant to the problem. Firstly, a "Loss" feature is added to the dataset. This feature is the difference between the "Theoretical Power" and "LV Active Power". Secondly, two features which represent the $x-component$ and $y-component$ of the wind speed and direction. The equation below highlights the calculation required for these two features.

$$X\_com = W_s.Cos(\frac{W_d.\pi}{180}) \tag{1}$$

$$Y\_com = W_s.Sin(\frac{W_d.\pi}{180}) \tag{2}$$

Where W_s and W_d are the wind speed (m/s) and wind direction (°) respectively.

Thirdly, a categorical feature is appended to the dataset. This feature represents the wind direction and is only used during the exploratory data analysis section of the project. It is removed prior to modelling.

Figures 1, 2 show the autocorrelation function (ACF) and partial autocorrelation function (PACF). The ACF shows how the current timestep has a correlation on the previous timestep. Figure 1 indicates a clear dependence with the timestep observed (t) and a time step of $t - 1$.

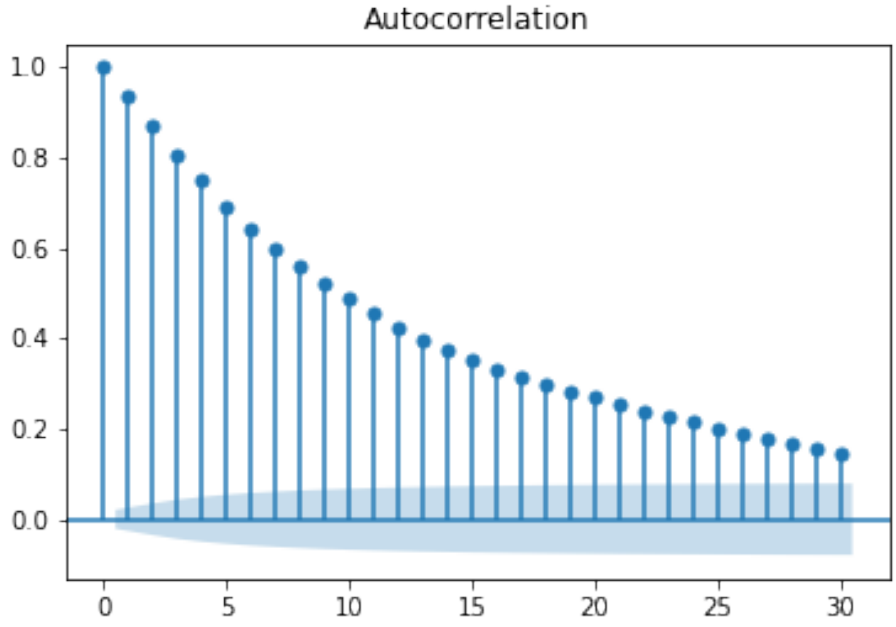

Figure 1: Autocorrelation function over 30 days.

Figure 2 shows that the PACF which represents the relationship between the observation and several timesteps before the observation, in this case 30 days. Figure 2 shows a time lag of one time step has the most significance. The last feature added to the dataset is a time lag of one (T_1) for the "LV Active Power" feature.

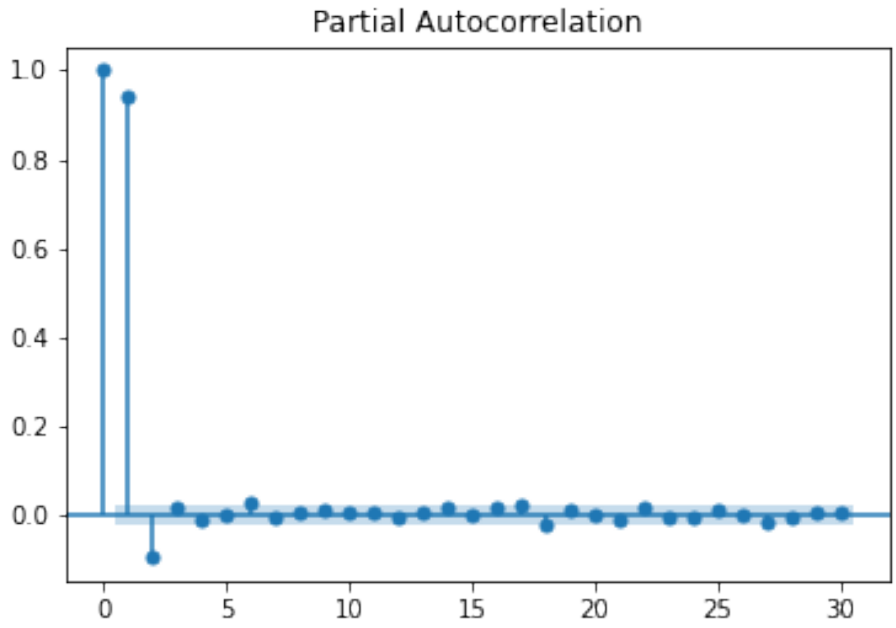

Figure 2: Partial autocorrelation function over 30 days.

3.3 Exploratory Data Analysis

This section will focus on exploring the data using visualisation tools, to have a better understanding of the different features and to justify the suitability for time series modelling.

3.3.1 Visualising the Data

Figure 3, 4 below show the distribution of "Active Power" and "Wind Speed" over the year. The active power shows peaks at $0kW$ and $\approx 3500kW$, which is understandable as energy from the wind is unpredictable.

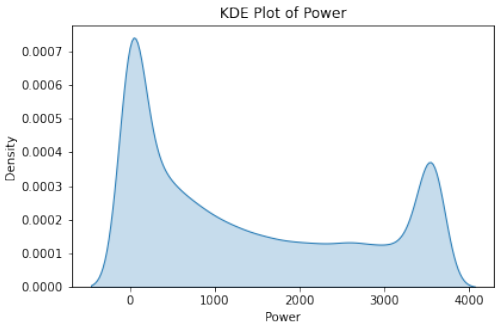

Figure 3: Kernal Density Estimate of Active Power.

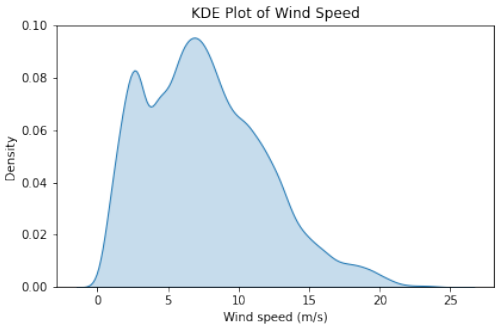

Figure 4: Kernal Density Estimate of Wind Speed.

Figure 5 below indicates the direction the wind was blowing throughout the year of 2018. It can be seen that North to East shows the direction the wind is most likely going to blow towards.

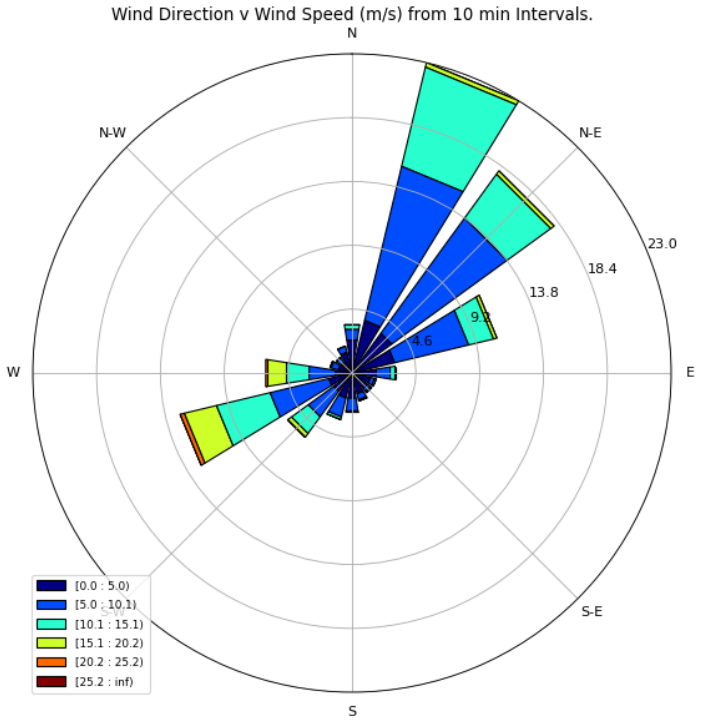

Figure 5: Windrose map of wind direction.

Examining the performance of the turbine can inform the operations and maintenance team on how efficiently the turbine is working. Figure 6 below shows how the actual power produced compares with the theoretical power as wind speed increases. The power produced by the turbine at wind speeds greater than $13m/s$ seems unsteady.

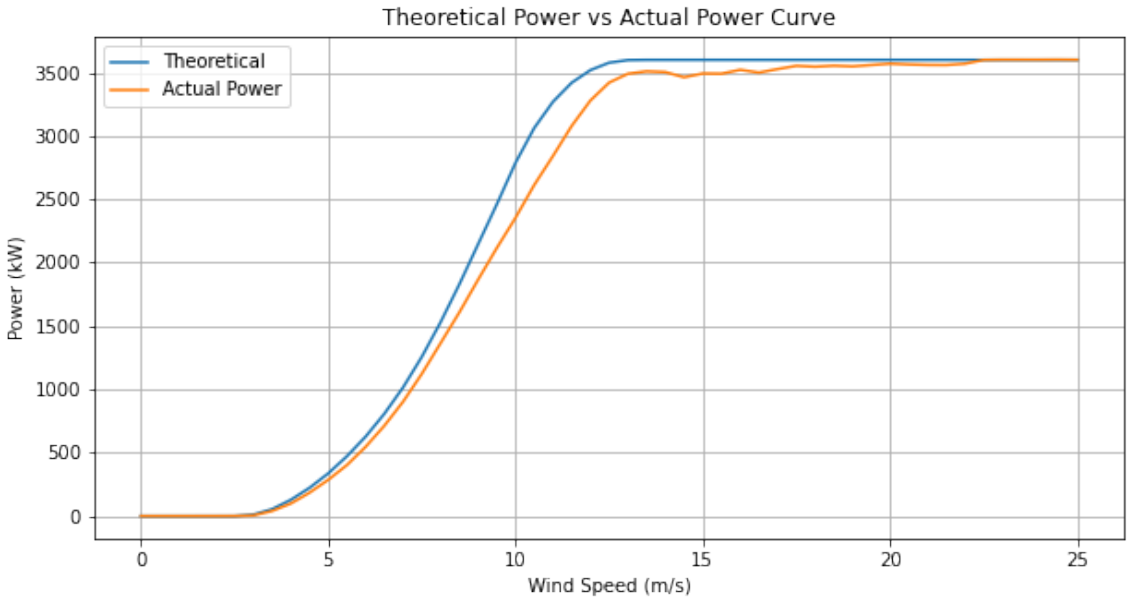

Figure 6: Comparison between the theoretical power and actual power.

Using the categorical direction feature of the dataset and the loss feature, we can investigate in which direction the turbine is not optimizing power performance. Figure 7 below identifies the ENE, NE, NNE and SSW directions as the most inefficient. Figure 8 shows a great decrease in performance in the NE and NNE direction at wind speeds grater than $13m/s$. This is worth investigating further and informing the operations and maintenance team to prevent further power loss in the future.

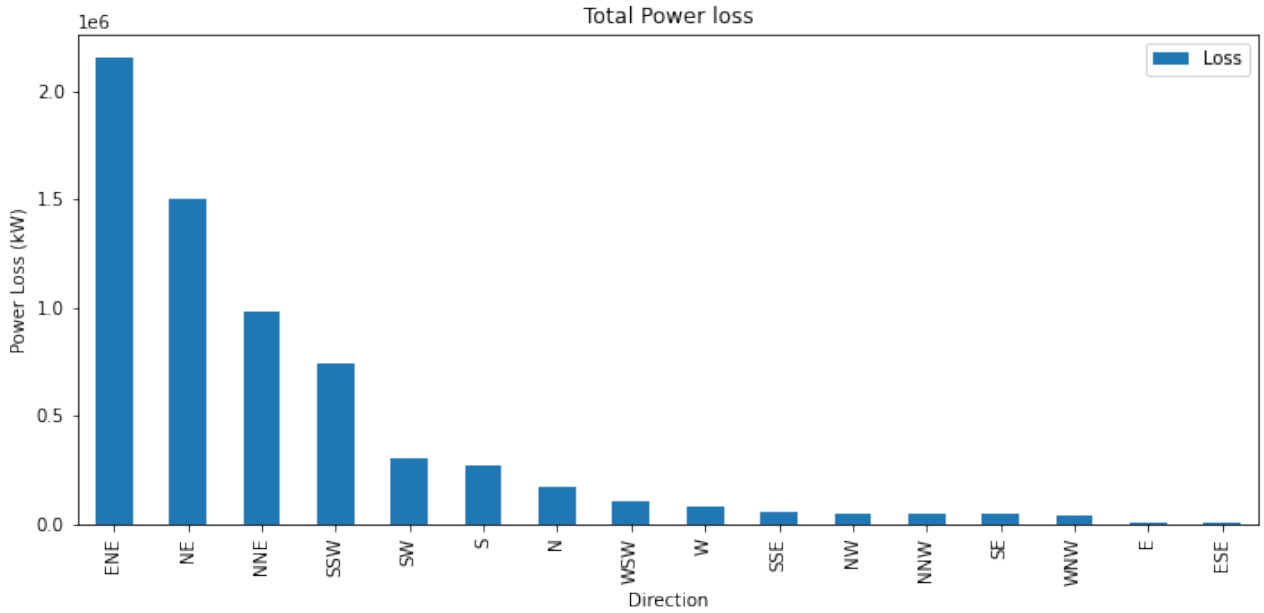

Figure 7: Power loss of the wind turbine in all categorical directions.

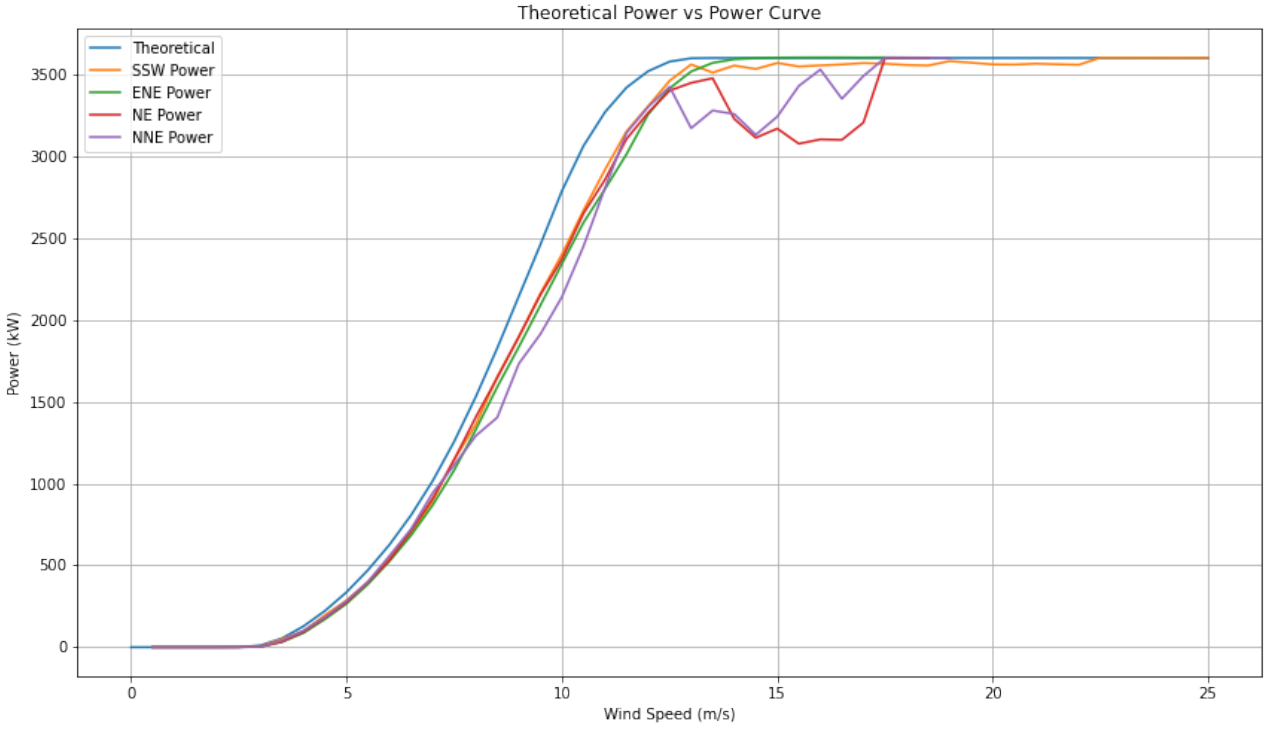

Figure 8: The four directions with the greatest power loss.

3.3.2 Co-integration between time series data

Investigating the relationship between features is a crucial for justifying the relationship between time series data. A hypothesis test is carried out to identify which features of the dataset are statistically significant.

H_o : If failed to be rejected, it suggests the LV power time series is not correlated with the respected time series feature.

H_1: The null hypothesis is rejected; it suggests the LV power time series is correlated with the respected time series feature.

For a $p - value < 0.05$, the null hypothesis is rejected and the two features are co-integrated. For a $p - value >= 0.05$ we fail to reject the null. Table 1 below showcases the $p - value$ for each feature in the dataset and shows all features are statistically significant.

Feature	Wind Speed	Theoretical Power	Loss	X_com	Y_com
p-value	3.177e-29	7.475e-28	9.449e-29	3.176e-29	3.178e-29

Table 1: P-values for all features of the dataset.

3.3.3 Stationarity test

In order to model time series data, the data needs to be stationary. Stationary means the statistical properties including the population mean (μ) and variance (σ^2) do not change over time. Stationary is important because many useful analytical tools and statistical models rely on it [1]. Using the Augmented Dickey Fuller (ADF) test, we can examine all the time series features. The hypothesis test below was carried out to determine if the features in the dataset are stationary.

H_o : If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.

H_1: The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.

Table 2 below shows the p-value from the ADF test. Similar to the co-integration test, if the $p-value >= 0.05$ we can fail to reject the null and if the $p-value < 0.05$ we can reject the null. All values from table 2 show a $p - value < 0.05$, confirming all the time series features are statistically significant and we can proceed to modelling.

Feature	Wind Speed	Theoretical Power	Loss	X_com	Y_com
p-value	4.828e-24	1.934e-29	1.084e-27	2.592e-23	3.203e-15

Table 2: P-values for all features of the dataset.

4 Modelling

This section will go into detail about the steps required prior to modelling and building a deep learning model. This includes splitting the data into training and testing, so the test data does not contaminate the model and creating datasets that are an appropriate shape for the model. A Long Short Term Memory (LSTM) artificial recurrent neural network is used to make predictions. By tuning the hyperparameters of the model the best model will be chosen based on suitable model metrics.

4.1 Training and Testing Data

To prevent the training data being contaminated with the test data, careful consideration needs to be taken to split the data. The first step is splitting the data into training and testing. As the data used in this project is from the year 2018, the training data includes all dates from $01/01/2018 - 31/11/2018$. December will be kept as the test data.

4.2 Prepossessing

For LSTM models, the data needs to be scaled between **-1 and 1** to meet the requirements of the default tanh function [2]. The features "Wind Speed", "Theoretical power", "Loss", "x_com", "y_com" and "T_1" are used as the predictor variables and "LV Active Power" as the target variable.

LSTM models need to have a three dimensional shape in the format [**Samples, timesteps, features**] [3]. The training and testing data is split into this format for modelling, where 2-weeks of the predictor variables will be used to make a single prediction of the target variable. A **walk forward validation** method is used to make hourly predictions for two weeks. The resulting dimensions of X_train, y_train, X_test, y_test are **(7679, 336, 6), (7679,), (408, 336, 6), (408,)** respectively.

4.3 LSTM Model

There are many hyperparameters to tune for an LSTM model. Some of these include the batch size, epochs, number of neurons, number of layers, dropout, the optimizer and optimizer learning rate. The chosen metric for this regression problem is the mean squared error (MSE) and is the metric used to compare the performance of the LSTM model on the validation data. The hyperparameters shown below were used for the final model.

- **Batch size**: 14
- **Epochs**: 35
- **Neurons**: 32
- **Layers**: 2
- **ADAM optimizer learning rate**: 0.0005
- **Dropout**: 0.05

4.3.1 Validation Data

Figure 9 below shows the convergence between the training and validation data. There is a convergence at ≈ 0.075. A validation split size of %30 is used for training the model.

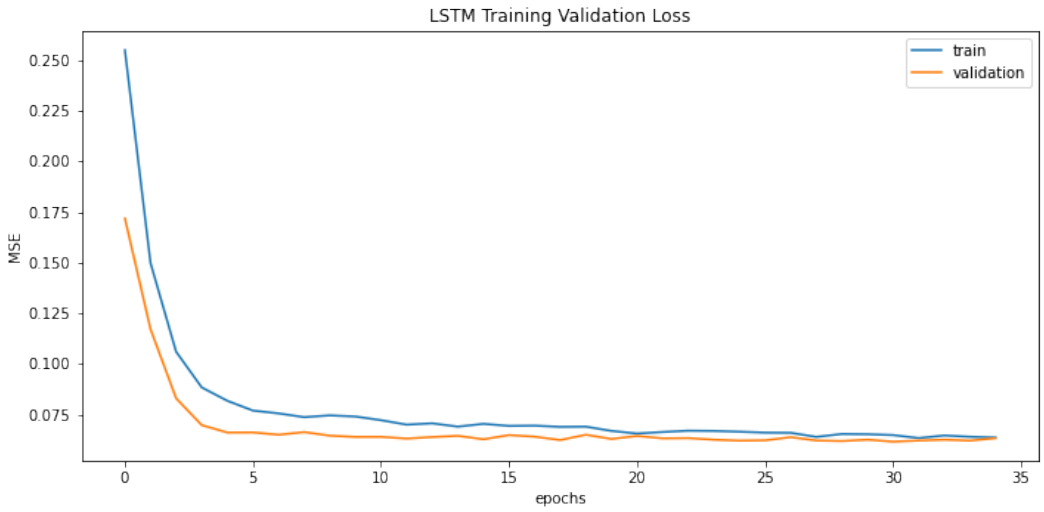

Figure 9: Converge between training and validation data.

4.3.2 Results

With the parameters shown above used for the final model, figure 10 shows the result of the LSTM on unseen data. The predicted values shown in blue compared to the real values, shown in orange. The model is able to predict the trend of LV Active Power and in this case able to predict the intermittency of the wind. The model is not able predict when the turbine will be at a standstill however can predict low power generation.

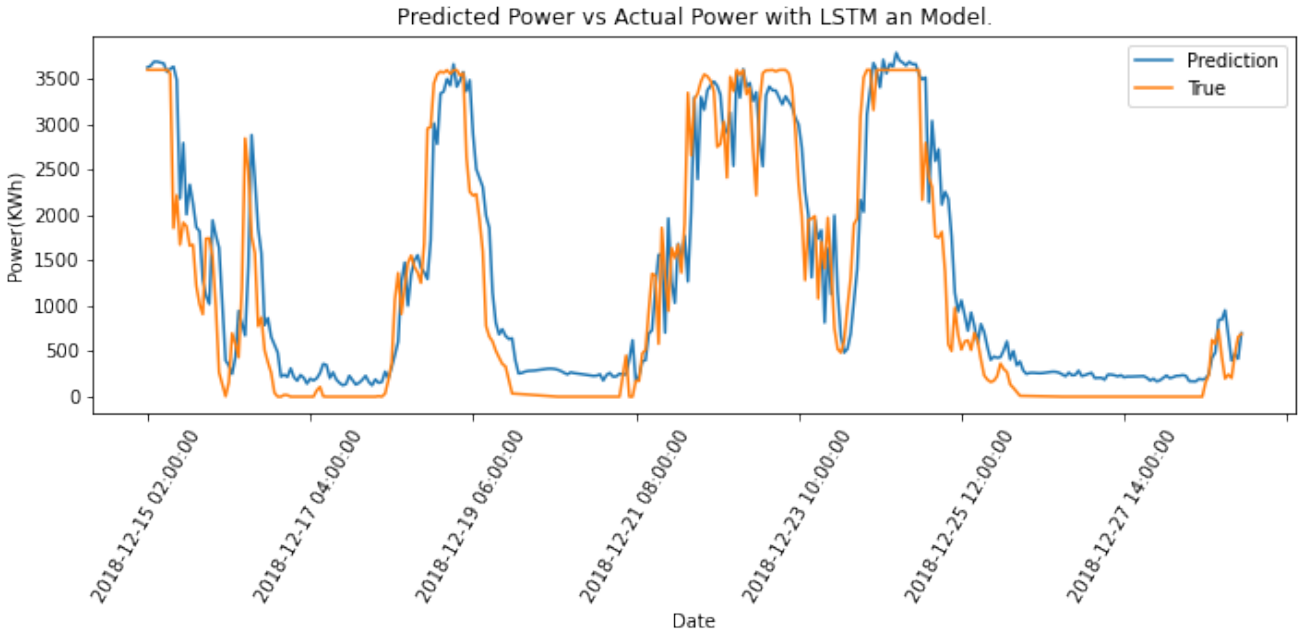

Figure 10: Comparison between LSTM prediction and actual test results.

The metrics of the final LSTM model are shown below.

- **RMSE**: 495.9
- **MAE**: 361.8
- $\mathbf{R}^2$: 0.87

5 Conclusion

This report identifies the necessary steps required to predict the power output of a wind turbine. The data used was of a wind turbine located in Turkey in the year of 2018. The data was resampled for hourly intervals and additional features were added including the wind speeds X and Y component and a time lag of one for LV Active Power. Further, relationships between the features were examined by use of exploratory data analysis, which concluded with high power loss in NE and NNE direction. Using the Augmented Dickey Fuller test, all features show co-integration with "LV Active Power" and tested to be stationary. The data was split into trianing and testing data, where the month of December was used for testing and the rest for training. A LSTM model was built, tuned and trained to find the optimum hyperparameters for the model. A final model was chosen with $\mathbf{RMSE} = 495.9$, $\mathbf{MAE} = 361.8$ and $R^2 = 0.87$. The model is able to predict the power output of the wind turbine on unseen data. This shows that an LSTM model is able to predict the intermittency of the wind and can be used as intelligence for the grid.

5.1 Assumptions & Limitation

The following assumptions and limitations need to be taken into consideration for this project:

- The Wind turbine is assumed to be fully functional after data filtering.
- The SCADA system provided accurate measurements throughout the year.
- A years worth of data is used.
- Additional wind data around the region could benefit the model.

5.2 Future Work

For anyone would like to continue the work carried out in this project, the following statements identify potential next steps:

- Predicting the direction the wind turbine should turn, to prevent power loss
- Obtaining relevant data (wind data, years prior to 2018) to improve model.

References

[1] Palachy S, 2019, Detecting stationarity in time series data, KD nuggets, URL: https://www.kdnuggets.com/2019/08/stationarity-time-series-data.html

[2] Brownlee J, 5^th August 2019, How to scale data for Long Short Term Memory in Python, URL : https://machinelearningmastery.com/how-to-scale-data-for-long-short-term-memory-networks-in-python

[3] Brownlee J, 17^th April 2017, How to Use Timesteps in LSTM Networks for Time Series Forecasting, URL : https://machinelearningmastery.com/use-timesteps-lstm-networks-time-series-forecasting/